

## **Towards Student Involvement in Essay Assessment**

**Aynur Yürekli  
Evrin Üstünlüoğlu**

İzmir University of Economics  
Balçova, Turkey

### **Abstract**

In language teaching, assessment is one of the most formidable challenges for both the students and the teachers. Especially, when the assessment of productive skills which are subjective by their nature are concerned, the "challenge" could very well turn into a "nightmare" for both parties. In order to avoid this undesired possibility, the attitude of the grader and the students towards the evaluation rubric is as vital as the rubric itself.

This study describes the standardization process of the writing rubric for the assessment of essays, which is accepted both by the graders and the learners who are subject to this evaluation. The paper outlines the phases of rubric revision and describes student involvement in essay evaluation. Special emphasis is put on how students used the rubric as a learning tool while writing their essays, and how they benefited from being familiarized to the rubric.

The results refer to the importance of inter-rater reliability, which is achieved by revising the assessment rubric in line with grader suggestions, and by checking consistency among graders of writing at certain intervals. The study also suggests that learner involvement in assessment promotes the outcome.

### **Introduction**

As interest in communicative competence in ESL/EFL continues to grow, more attention has been given to students' productive skills—speaking and writing and to their assessment because a lack in these skills immediately affects the message you are trying to communicate. Speaking involves pronunciation, intonation, accuracy and fluency; writing, which is more complex, requires knowledge of formal structure, style and punctuation. The subjective nature of the assessment of these skills means it is a formidable challenge, and therefore, needs more attention.

Writing has been the center of attention in ELT world for years and has engaged numerous researchers, particularly in terms of assessment. Literature indicates that the assessment process for writing is complex and may cause some problems, especially in essay evaluation. Shaw (2001) highlights three important issues about the assessment of second language writing ability:

- Features which distinguish second language writing performance at different levels of proficiency,
- Process by which writing examiners form judgments about scripts,
- Achievement of an acceptable level of reliability in written assessment.

Diederich, French and Carlson (1961), and Vaughan (1991) state that different markers respond to different facets of writing and focus on different essay elements and perhaps have individual approaches to reading essays. This problem becomes more striking, particularly when faculty teachers share multi-section courses. Studies highlighting the grader as an important variable indicate several factors affecting assessment such as teacher attitude, gender, experience and professional background, and the type of the assessment they use- holistic or analytical (Boughey, 1997). Teacher attitude becomes very important when grades vary depending upon how lenient or strict a student's grader is. In addition to this, Freedman (1979) states that graders are inconsistent in terms of what features to take as important. Variables such as gender, professional background, amount of exposure to L2 writing have also an impact on assessment (Hamp-Lyons, 1990; Vann, Lorenz and Meyer, 1991). Finally, the kind of assessment criteria used affects judgment.

Studies indicate not only the importance of graders but also that of students in the assessment process. Stiggins (2001) argues that students are "the key assessment users" (p.17) and should be able to use assessments in similar ways that teachers use them. To inform the students about the assessment process and to enable them to succeed, teachers mostly prefer using rubrics as a means of communicating expectations for an assignment, providing focused feedback on works in progress, and grading (Andrade, 2000; Goodrich, 1997). Particularly, when used as part of a formative, student-centered approach to assessment, rubrics have the potential to help students make judgments about the quality of their own work (Stiggins, 2001, p.11). In his research, Andrade (2000) concluded that explaining a rubric to students was associated with higher scores as students had the opportunity to understand how their writing was evaluated and the qualities of effective writing, as defined by the rubrics they received. In recent research, Hafner and Hafner (2003) discovered a high correlation between student and instructor ratings, providing evidence that undergraduate students can be effective users of rubrics.

Due to the grader and student impact in writing assessment, the studies in the field recommend that investigating the processes rater judgement is one way to reach a greater understanding of rater behavior (Hamp-Lyons, 1990; Milanovic, Saville and Shuhong, 1996). One way of investigating this process is to organize standardization or norming sessions together with training sessions. Known as "norming" or "inter-rater reliability", training sessions seem to yield positive results.

This study attempts to explain how training and norming sessions of graders, together with student engagement in assessment, promote achievement and stimulate learning behaviors typically associated with academic success.

### **The aim of study**

The aim of this study is, first, to account for the inter-rater reliability with the criteria used as an assessment tool for essay evaluation. Next, it aims at testing whether students' participation in the assessment process makes a difference in the evaluations.

### **Methodology**

The Methodology section covers the description of the institution, participant teachers and students, and the data collection instruments.

#### ***Scope of the Study***

##### **The Institution**

The study was conducted at İzmir University of Economics, School of Foreign Languages. The School of Foreign Languages offers three different programmes, one of which is the Freshman Academic Reading and Writing Skills. This course aims at teaching the basics of academic reading and writing. As this study focuses on the assessment of writing, essays in particular, the scope is limited to the teachers and students at Freshman English programme.

##### **Participant Teachers**

All teachers participating in this study ( $n_1=20$ ), consisting of eleven native speakers and nine non-native speakers, were teaching Academic Reading and Writing Skills to all freshman students enrolled at İzmir University of Economics. Eleven of these teachers were native and nine were non-native. The teachers had more than five years of experience as English language instructors. The participant teachers provided data on the rubric's inter-rater reliability studies and the final version of the rubric.

##### **Participant Students (The Sample Group)**

In the School of Foreign Languages, 1275 students were enrolled in the Academic Reading and Writing Skills course when the study was conducted. As this study represents student initiation into the assessment of productive skills, 18 ( $n_2=18$ ) one section of students was chosen at random ( $n_2=18$ ) to be the sample group, referred to throughout this study. All of these students were taking the course for the first time. The sample group provided the study with the data on the essays of phase I and phase II.

#### ***Data Collection Instruments***

The data of this study were collected in five phases which were carefully planned and implemented parallel with a time-table.

## Sample Essays

To check the consistency among the graders of the final writing exam, three essays were used. These essays were chosen according to their original scores; one essay was taken for each category on the basis of the following scores: 0-50-low, 51-70-average, 71-100-high.

### Rubric-1st Version

In this study, rubric is defined as “a document that articulates the expectations for an assignment by listing the criteria, or what counts, and describing levels of quality from excellent to poor”. (Andrade, 2000).

A rubric for essay development had been adapted in the previous years. However, the pilot study with the teachers on inter-rater reliability revealed that this required some changes and the teachers needed some standardization sessions to ensure consistency among graders. The observed difference in the scores that the teachers gave for the three essays ranged from 42-46. To see whether this difference was statistically significant or not, ANOVA was used. The scores given by the graders is given in Appendix A.

### Rubric-Final Version

After having calculated the inter-rater values of the 1st version of the rubric, the problematic areas of the rubric were identified and improved. Moreover, some issues regarding the layout and user-friendliness of the rubric were changed, based on teacher-grader feedback. The final version of the rubric consisted of four components: Structure (30%), Content (30%), Vocabulary (20%) and Language (20%). “Structure” covered the organisational elements of the essay such as the thesis statement, topic sentences, supporting ideas and concluding paragraph; “Content” covered the features of relevance, progression of ideas and clarity; “Vocabulary” covered the use of correct and varied vocabulary with special emphasis on the use of vocabulary taught during the class; and “Language” covered the use of accurate language and punctuation.

### Essays of Phase I

This phase served as the pre-test phase of this study. To see whether the quality of students’ essays differs after being introduced to the criteria and being familiarized with the criteria, the students (n=18) were asked to write a 5-paragraph essay on “the advantages of internet” without knowing the assessment criteria. The collected essays were evaluated by both researchers separately by using the final version of the rubric and the average of the gradings was taken as the essay score. No feedback was given to the students on the weaknesses and strengths of their essays, neither were the scores announced.

### Students' evaluations

In order to familiarize the students with the rubric that the graders used, students were given the same three essays that the teachers initially used in the standardization session and were asked to evaluate the essays using the criteria. This session was conducted in class with the researchers so that students had the chance to ask for clarification regarding the rubric as needed. The aim was to see whether there was consistency among students in understanding the rubric, and thus the expectations of those who would later grade their papers (For student scores, see Appendix B).

### Second Essays

This phase of the study served as the post-test for student essays. The student sample group was again asked to write a 5-paragraph essay on “the advantages of internet”. The purpose was to see whether there was any significant improvement in student essays after they familiarized themselves with the rubric and used it to grade essays. The essays were evaluated by both researchers separately by using the final version of the rubric and the average of the gradings was taken as the score of the essay. The t-test was computed to see whether there was a significant difference between the students' scores on the first essay and the scores on the second essay.

## Data Analysis and Findings

### Teacher Grading Results

The three essays written by different students were graded by 20 teachers using the first rubric prepared by the School of Foreign Languages testing committee. The scores that the teachers assigned to the individual essays referred to a score difference of 46 when lowest and highest grades were taken into consideration (see Appendix A). The results were subjected to ANOVA analysis to see whether graders were consistent among themselves in their gradings. Table 1 shows the results of the ANOVA test.

Table 1. Results of the Gradings with Rubric 1-Teachers

S					
	a	d		F	
0	0	9			
3	3	0	8		
3	3	9			

As the results for “between groups” show, ( $p > .05$ ) the scores assigned to the essays with the first rubric refer to a low consistency among teachers. Based on these results, the second rubric was developed and a second assessment with the new rubric was carried out using the same essays.

After having developed the second rubric by taking teachers' views and feedback into consideration, the same procedure was duplicated with the second rubric. The results of the scores given with the second rubric showed a difference of 20 points among teachers (see Appendix A). To see whether this difference is statistically significant or not, ANOVA was carried out (see Table 2).

Table 2. Results of the Gradings with Rubric 2-Teachers

S		f		F	
W	9	9	6	3	
W	0	0			
W	3	9			

The grading results of the second rubric ( $p > .05$ ) reveal that the discrepancy among the graders had significantly dropped when the second rubric was used while grading the same essays. It is seen that the second rubric brought graders closer in their evaluations. Thus, it was chosen as the rubric for the assessment tool of essays.

*Student Results*

Essay 1

The first essays, which were written by the students without any familiarization to the rubric, were graded by both researchers and the average of these scores were assigned to the essays. The scores of both the first and second essays are given in Table 4.

*Student Gradings*

The students, after having been familiarized with the rubric, were asked to grade the same set of essays as the teacher graders did, to see whether they had a similar perspective with the teachers in the assessments of their essays (see Table 3). To see whether the students gradings were comparable to the ones given by teachers, t-test was used.

Table 3. Results of the Gradings with Rubric 2-Students

S		f		F	
W	8	7	6	4	
W	0	0			
W	3	3			

The results indicate that students are quite consistent among themselves in their grading, thus their understandings of what constitutes a good essay. The scores given by the students are consistent with the scores given by the teachers. The ANOVA results show that the students and teachers share the same perspectives ( $p > .05$ )

## Essay 2

After being familiarized with the criteria and having used it to evaluate the sample essays, students were asked to write an essay on the same topic as the previous one “advantages of internet”. The purpose was to see whether there was a positive change in the quality of their essays after being introduced to the rubric. The scores and means of the scores assigned by two independent graders are given below:

Table 4. Student Results of Essay 1 and Essay 2

<i>Students</i>	<i>Essay-1</i>	<i>Essay-2</i>
Student-1	65	78
Student-2	42	65
Student-3	48	44
Student-4	70	82
Student-5	69	79
Student-6	54	65
Student-7	49	60
Student-8	63	75
Student-9	71	79
Student-10	57	68
Student-11	89	93
Student-12	65	78
Student-13	57	54
Student-14	48	43
Student-15	63	75
Student-16	72	82
Student-17	66	75
Student-18	53	69
<i>Mean</i>	<i>61.16</i>	<i>70.22</i>

T-test was computed to see whether there was a meaningful difference between the scores that students got in essay 1 and essay 2. The results are shown in Table 5.

Table 5 T-test Results of Essay 1 and Essay 2

		Pre	Post
Pre	M	1	*
	S	,	
	N	8	8
Post	M	*	1
	S	0	,
	N	8	8

\* p < .05

The results suggest that there is a significant difference between the scores of Essay 1 and Essay 2 ( $r=.846$ ), which means that there is an improvement in students' essay after they were introduced and familiarized with the rubric.

### Discussion

The results of this study suggest that teacher involvement in rubric preparation clarifies the components to be considered while grading and the standardization sessions conducted afterwards improve essay assessment, especially in terms of inter-rater reliability. The results also reveal that student involvement in the evaluation process helps students to understand what constitutes a good essay, and thus has a positive impact on the essays they write.

Boughley (1997) highlights that each grader has distinct experience, perspective, personality and skills, which inevitably leads to differences in grading. The results of this study show that when teachers' views are taken into consideration and a rubric developed accordingly, the inter-rater reliability really improves. This finding supports Freedman (1979), who states that graders might be inconsistent in terms of which features are considered as important. The results of this study agree with the literature that standardization sessions among teachers improve essay assessment because they contribute to the consistency among graders (Hamp-Lyons, 1990; Weigle, 1994; Milanovic, Saville and Shuhong, 1996).

Literature indicates that not only the graders but also the ones who are graded are important in the assessment process. As such, students need to receive a training in a similar way the graders do. Thus, students will be informed about the process of grading and what is exactly expected from them. The findings of this study have demonstrated that students' knowledge about the rubric and assessment procedure makes a meaningful difference in students' productions. These results are consistent with the claims of Stiggins, (2001), Andrade (2000) and Goodrich (1997). Furthermore, the results suggest that student gradings are quite similar to the gradings of the teachers, consistent with the conclusions reached by Hafner and Hafner (2003).

### Implications and Conclusion

This study foregrounds the importance of inter-rater reliability, which can be achieved by revising the assessment rubric in line with grader suggestions, and by checking consistency among graders of writing at certain intervals. Furthermore, the study also suggests that learner involvement in assessment promotes the outcome. Making students well-informed about the assessment tool and giving them a chance to actively use the rubric helps them to achieve higher level of success.

In conclusion, it can be said that assessment involves two parties, the teachers as well as the students. Learner involvement does not only shed light on the evaluation process, but also has a direct and positive impact on the student outcome, as far as essay writing is concerned.

### References

- Andrade, H. 2000: Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13-18.
- Bachman, L.F. 2000: Modern language testing at the turn of the century: Assuring that what we count counts, *Language Testing* 17/1, 1-42.
- Bougey, C., 1997: Learning to write by writing to learn: a group-work approach, *ELT Journal* Volume 51/2, April 1997. Oxford University Press.
- Diederich, P.B.; French, J.W and Carlton, S.T. 1961: Factors in judgements of writing ability. *ETS Research Bulletin RB-61-15*, Princeton N J: Educational Testing Service.
- Freedman, S W. 1979: Why do teachers give the grades they do? *College Composition*, 32, 365-387.
- Goodrich, H. 1997: Understanding rubrics. *Educational Leadership*, 54(4), 14-17.
- Hafner, J., & Hafner, P. 2003: Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25 (12), 1509-1528.
- Hamp-Lyons, L. 1990: Second Language Writing: assessment issues. In B Kroll (Ed) *Second Language Writing. Research Insights for the Classroom*, Cambridge University Press.
- Milanovic, M.; Saville, N. And Shuhong, S. 1996: A study of the decision-making behaviour of composition markers. In M. Milanovic and N. Saville (Eds), *Studies in Language Testing-Performance Testing, Cognition and assessment:*

Selected Papers from the 15th Language Testing research Colloquium.  
University of Cambridge Local Examinations Syndicate and Cambridge  
University Press.

Stiggins, R.J. 2001: Student-involved classroom assessment (3rd ed.) Upper Saddle  
River, NJ: Merrill/Prentice-Hall.

Vann, R.J., Lorenz, F.O and Meyer, D.M. 1991: Error gravity: faculty response to  
errors in the written discourse of nonnative speakers of English. In L Hamp-  
Lyons (Ed), *Assessing Second Language Writing in Academic Contexts*.

Vaughan, C. 1991: Holistic assessment: what goes on in the rater's mind? In L. In  
L Hamp-Lyons (Ed), *Assessing Second Language Writing in Academic  
Contexts*.

Weigle, S.C 1994: Effects of training on raters of ESL compositions, *Language  
Testing*, 11/2, 197-223.

Shaw, D.S. 2001: Issues in the assessment second language writing. *Research  
Notes*. University of Cambridge Local Examinations Syndicate and Cambridge  
University Press.